Internet Protocol Suite

Introduction

CSCI E 45a: The Cyber World – part A

Introduction: learning goals

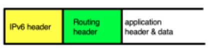
- Understand the architecture and functioning of the Internet Protocol, both version 4 and version 6
- Understand how IPv6 was developed and why
- Understand the functioning and use of the basic set of higher-level IP protocols

Introduction: this module

IPng

- This module mostly covers technology but includes some history

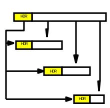
Introduction: topics

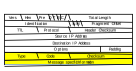


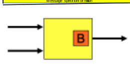
- Addresses – R
  IP internetwork address format and allocation
- Internet Protocol – R
  IP packet format and operation
- IP Next Generation – O
  The IETF's IP next generation effort
- IPv6 headers – R
  The IPv6 optional headers

4 Copyright © Scott Bradner & Ben Gaucherin 2015

Introduction: topics, contd.



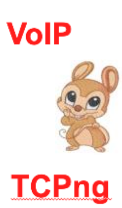




- Fragmentation – R
  IP packet fragmentation
- Riding on IP – R
  Layered encapsulation
- ICMP - R
  Internet Control Message Protocol
- Flow and congestion – R
  The difference between flow and congestion control

5 Copyright © Scott Bradner & Ben Gaucherin 2015

Introduction: topics, contd.



- UDP – R
  User datagram protocol
- TCP – R
  Transmission Control Protocol
- QUIC - R
  QUIC

6 Copyright © Scott Bradner & Ben Gaucherin 2021

Image credits

All drawings and photos by Scott Bradner unless noted

Slide#	credit
2	IPv6 logo – IPv6 Forum http://www.ipv6forum.com/ipv6_enabled/approval_list.php
4	IETF logo – IETF – https://www.ietf.org/logo/ietf-logo.gif
5	hourglass –Jonathan Zittrain - e.g. http://dltj.org/article/hourglass-national-e-book-program/
6	mouse clip art – http://embroiderypassbook.com/index.php?main_page=product_info&cPath=4&products_id=1166

Internet Protocol Suite

Internet Addresses

CSCI E 45a: The Cyber World – part A

1

Internetwork (IP) addresses

- Two part address field
  - Network identifier + host identifier
    - Host is within the local network
- IPv4: 32-bit address field
  - 4,294,967,295
- IPv6: 128-bit address field
  - 340,282,366,920,938,463,463,374,607,431,768,211,456
- Boundary between parts
  - IPv4: Configured
  - IPv6: Generally 64 bits
- Identifies a network interface
  - Interfaces can have >1 address

2

IP address, functions

- A locator
  - Where this host is in the internetwork
- An identifier
  - Which host is this host
- Might be useful to separate functions

3

Separate ID & Locator


Bob Moskowitz

- If ID&L are the same thing then mobility an issue
  - If you change locations while communicating, TCP breaks
    - Because higher-level protocols use full IP addresses in checksum
- Multiple Endpoint IDentifier (EID) proposals
  - See (e.g.,) Host Identity Protocol (HIP)

4

EID issues

Identifier

Locator

- Security: having combined means spoofing is harder
  - The routing system will not forward packet to the "wrong place"
    - But this is only meaningful for two-way conversations
    - No way to be sure where a packet came from
- Management: hard to map IDs to locators (big database)
  - Might be good for privacy

5

Representing addresses

128.103.8.36

- IPv4: "dotted quad"
  - 4 decimal values (one per byte) separated by periods

1080::8:800:200C:417A
0::0

- IPv6: hex string
  - Suppression of string of contiguous zeros - use "::"

128.103/16

- Often need to indicate the network part of address
  - Called "prefix"
    - /16 means 16 bits of prefix

6

Localhost & Loopback addresses

IPv4: 127.0.0.1
IPv6: ::1

- Localhost address
  - Address that always means "this host"
- Loopback address
  - Address assigned to the host rather than to a network interface
  - More reliable way to address a function such as node management
    - Survives as long as there is any interface working

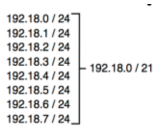
Address assignment

- Top level: IANA (part of ICANN)
  - Allocate big blocks of addresses (address prefixes) to Regional Internet Registries (RIRs)
    - 5 RIRs, each with own geographic territory
- RIRs allocate smaller address prefixes to ISPs
  - And to some multi-homed end sites
- ISPs allocate address prefixes to customers
  - Some customers can be smaller ISPs

Classful & classless addresses

- IPv4 used to have "classful addressing"
  - Class A, B, C, D & E
  - Defined large blocks of address space
    - e.g., Class B = 65,535 addresses
  - Dropped in 1994 to increase assignment efficiency
- Now use "classless addressing"
  - Assignment block size defined by "prefix length" in bits
    - e.g., 128.103/16 = 65,535 addresses

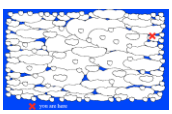
Address aggregation (CIDR)

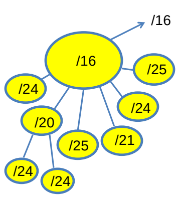


- Adjacent bocks of addresses can be aggregated into a shorter prefix
- Classless InterDomain Routing
- Classless addresses can be hierarchically assigned
  e.g., an ISP is allocated a /16
  - Assigns some /27s and /25s (etc.) out of the /16 to customers
  - Advertises whole /16 to the rest of the Internet

10 Copyright © Scott Bradner & Ben Gaucherin 2015

Hierarchical Routing and Addressing

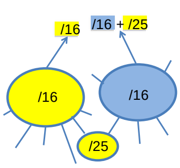




- Internet topology is a rough hierarchy
  ISPs and their customers
  - ISPs can also be customers of other ISPs
- Physical topology hierarchy must be reflected in address assignment to permit aggregation
  - w/o aggregation routing tables would have to include all the individual networks that make up the Internet

11 Copyright © Scott Bradner & Ben Gaucherin 2015

CIDR Issues



- New customers must renumber to provider's space
  ISPs require renumbering to save money, no regulations
- Tends to bind customer to provider
  ISP retains rights to addresses
- Problem with sites multi-homed to > 1 ISP
  2nd ISP must inject an exception into the routing table

12 Copyright © Scott Bradner & Ben Gaucherin 2015

Private addresses

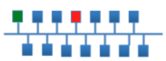
10/8

172.16/12

192.168/16

- RFC 1918: *Address Allocation for Private Internets* set aside some IPv4 addresses for use in private networks
- Must not be routed in Internet
- Originally for nets not connected directly to Internet
- Now also used when using NATs or firewalls which do address translation
  - E.g. WiFi access points

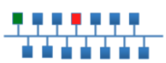
Address types



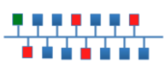
Unicast: a single destination

Must be unique within network scope

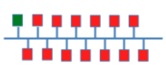
- Global: IPv4 & IPv6
- Private: IPv4
- Link-local: IPv6



Anycast: topologically closest node – IPv4 & IPv6

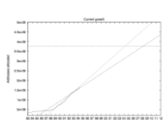


Multicast: nodes subscribed to a group – IPv4 & IPv6

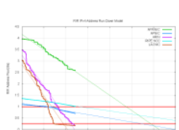


Broadcast: all nodes on a LAN – IPv4 only

IPv4 address end game



1994 projection



IPv4 Address Runout

- Now actually running out of IPv4 addresses
  - (educated) guess in 1994 & 1995: 2008 ± 3
- IANA ran out 3 Feb 2011
  - RIRs out or running out
- Now a market in IPv4 addresses
  - Migration to IPv6 seen as "harder"
- Hierarchical address assignment lost in a market

Image credits

All drawings and photos by Scott Bradner unless noted

Slide # credit

4 Moskowitz photo: iot-360.eu/2014/speaker/robert-moskowitz/

8 Logos: IANA (www.iana.org), ICANN (www.iacann.org), ARIN (www.arin.net), RIPE NCC (https://www.ripe.net/) APNIC (www.apnic.net), LACNIC (www.lacnic.org) & AFRINIC (www.afrinic.net)

15 top chart - Tony Li 1994 - http://www.ietf.org/proceedings/30/ipng/ale.html Slides - Li

bottom chart - http://www.potaroo.net/tools/ipv4/ - June 15 2015

16

Internet Protocol Suite

Internet Protocol

CSCI E 45a: The Cyber World – part A

Internet Protocol (IP)

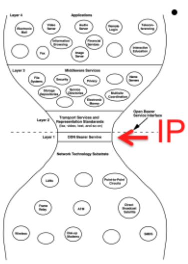


FIGURE 2.1 A four-layer model for the Open Data Network

- Datagram-based bearer service
  - Self contained
  - Handled independently of preceding or following packets
  - May contain processing hints
  - No delivery guarantees
    - Net may drop, duplicate, & deliver out of order
    - Reliability (where needed) must be done at higher levels
  - Contains destination and source internetwork addresses

IP, contd.

Jon Postel

- Designed to deal with a network of networks
- To get a packet to the correct node on the correct net
- Scope of IP

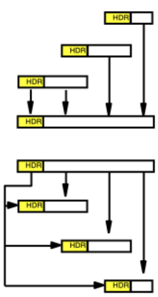
  specifically limited in scope to provide the functions necessary to deliver a package of bits (an internet datagram) from a source to a destination over an interconnected system of networks

  RFC 791

IP, features

- Datagram
- Best-effort
- No delivery guarantees
- No delivery-order guarantees
- No session-based state required in network
  - But may be present
    - e.g., NAT & firewall
- Can run over many types of networks

IP's jobs



1/ Must choose "next-hop" on path to destination

2/ Must be able to reassemble fragmented datagrams

Only at destination host

3/ May fragment datagrams that are too large for part of the path

source hosts (IPv4 & IPv6)

routers on path (IPv4)

4/ Must provide diagnostic and error functionality

RFC 1122 - sec 3.1

Robustness Principle


Jon Postel

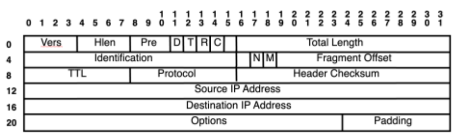
"Be conservative in what you do, be liberal in what you accept from others."

RFC 793

Internet Protocol

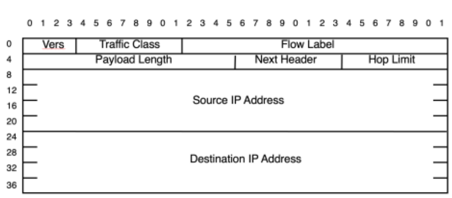
Internet Protocol Headers

IPv4 header



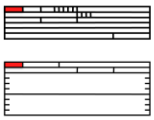
- Fixed length base header
- Variable number of options
- Padded to 32 bit word alignment

IPv6 header



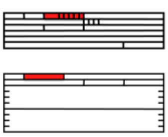
- Fixed length header
- No header options
- 64-bit word alignment

IP header: version number



- Identifies protocol version #
- 1-3: development versions of original protocol
- 4: version used on the Internet
- 5: Internet Stream Protocol (ST)
- 6: IP next generation

IP header: difserv/CE, Traffic Class field



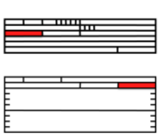
IPv4: updated to replace multiple fields with a difserv/CE field

Differentiated services – 6-bits

Congestion experienced – 2-bits

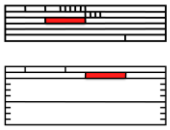
- IPv6: same as new IPv4

IP header: loop detection



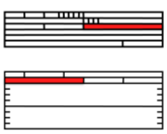
- IPv4: Time To Live
  - Decremented by each hop & if hop delays packet by 1 sec
- IPv6: Hop Limit
  - Decremented by each hop
- Discard packet if decremented to 0
  - Unless packet destination address is that of the local host
- Return ICMP message if packet discarded

IP header: Protocol/Next Header



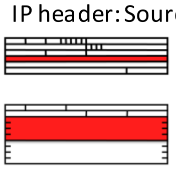
- IPv4: Protocol
- IPv6: Next Header
- Specifies what is next in the datagram

IP header: Total /Payload Length



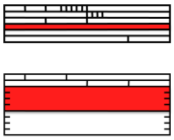
- IPv4: Total Length
  Length of packet, including header
  Note: header is variable length
- IPv6: Payload Length
  Length of payload (not including header)
  Note, header is fixed length

IP header: Source Address



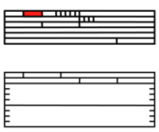
- IP address of sending node
- IPv4: 32 bits
- IPv6: 128 bits

IP header: Destination Address



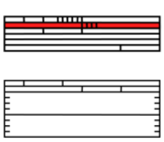
- IP address of destination node
- IPv4: 32 bits
- IPv6: 128 bits

IPv4-only header fields: Header Length



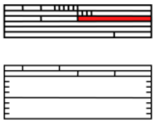
- Length of IP header
- Not needed in IPv6 because IPv6 has a fixed length header

IPv4-only header fields: Fragmentation



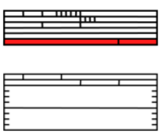
- Multiple fields to support fragmentation
- In-route routers do not fragment in IPv6 so header not needed in every packet
  - Reduce demands on routers
- Separate fragmentation header in IPv6, only used if sending host fragments packet

IPv4-only header fields: Header Checksum



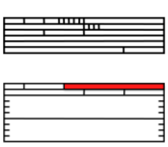
- Checksum that covers just the header
- Not used in IPv6 – determined that the downside of corrupted headers is less than adding processing to the routers

IPv4-only header fields: Options & Padding



- IP options – e.g.:
  - Strict Source Route
  - Lose Source Route
  - Record route
  - Time stamp
  - Traceroute
  - Router Alert
- Options use their own header in IPv6

IPv6-only header field: Flow Label



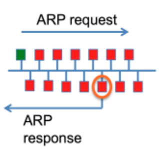
- Identifies a packet flow
  - A series of packets between an application in one host to an application in another host
    - Specifically, packets with the same source & destination addresses & port values and protocol value
- Can be used by a router as a request to treat a series of packets the same way
  - Reduces the chance of reordering
- Currently generally ignored

Internet Protocol

Finding Neighboring Nodes

Finding Neighboring Nodes

- Need LAN Media Access Control (MAC) address to get packet to correct node on LAN
- IPv4: Address Resolution Protocol (ARP)
  - Nodes & routers
    - Send broadcast ARP query that includes target IP address
    - Node with target address responds with ARP response that includes its MAC address



Finding Neighboring Nodes, contd.

- IPv6 - Neighbor Discovery (ND)
  - Nodes
    - Send multicast ND query with including target IP address to "selected node multicast address"
    - Node with target address responds with a ND response that includes its MAC address
  - Routers
    - Routers advertise themselves and their MAC addresses with Router Announcements

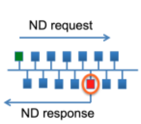


Image credits

All drawings and photos by Scott Bradner unless noted

Slide# credit

2 Hourglass - Realizing the Information Future - http://www.nap.edu/openbook.php?isbn=0309050448

3 Postel photo - http://www.wired.com/2012/10/joe-postel/

6 Postel photo - http://www.wired.com/2012/10/joe-postel/

Internet Protocol

IPng

CSCI E 45a: The Cyber World – part A

IP Next Generation

I E T F

- How did IPv6 come about?
- The reason and the process

As it was in 1990

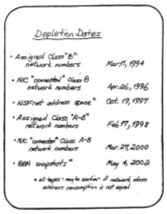
A: 8 bits net, 24 bits host

B: 16 bits net, 16 bits host

C: 24 bits net, 8 bits host

- Classful IP address assignment
  - Very inefficient allocation
    - A: 16,777,216 addresses
    - B: 65,536 addresses
    - C: 256 addresses
- Assignments made to end sites
- Internet was growing, class B was the common assignment size

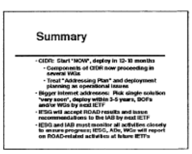
July 1990, Vancouver IETF meeting



- Frank Solensky reviewed the IP assignment statistics
- Determined that Class B addresses would run out in mid 1994 at the rate they were being assigned
  - The press: the Internet is running out of all addresses
- The IAB formed the Routing and Addressing (ROAD) special working group

4

March 1992, ROAD report


Summary

- CIDR: Start "NOW", deploy in 12-18 months
  - Components of CIDR now proceeding in several WGs
  - Treat "Addressing Plan" and deployment planning as operational issues
- Bigger Internet addresses: Pick single solution "very soon", deploy within 3-5 years, BOFs and/or WGs by next IETF
- IESG will accept ROAD results and issue recommendations to the IAB by next IETF
- IESG and IAB must monitor all activities closely to ensure progress; IESG, ADs, WGs will report on ROAD-related activities at future IETFs

- ROAD working group recommendations:
  - Switch to classless address assignments & processing
  - Pick a design for a IP next generation that supported bigger addresses
- IETF created new temporary IPng area in July 1993
  - Moved all related WGs to new area

5

IPng Area


Scott Bradner


Allison Mankin

- Assign two current area directors as IPng ADs
  - Scott Bradner (OPS)
  - Allison Mankin (TSV)
- 3 proposal working groups
  - SIPP: Simple Internet Protocol Plus
  - TUBA: TCP and UDP with Bigger Addresses
  - CATNIP: Common Architecture for Next Generation Internet Protocol

6

IPng Area, contd.

J. Allard - Microsoft
Steve Bellovin - AT&T
Jim Bound - Digital
Ross Callon - Wellfleet
Brian Carpenter - CERN
Dave Clark - MIT
John Curran - NEARNET
Steve Deering - Xerox
Dino Farinacci - Cisco
Paul Francis - NTT
Eric Fleischmann - Boeing
Mark Knopper - Ameritech
Greg Minshall - Novell
Rob Ullmann - Lotus
Lixia Zhang - Xerox

- Appointed a directorate
- Formed Address Lifetime Expectations WG
  - Estimate: address run out 2008±3
- Other working groups for transition, autoconfiguration, testing, etc.

7

IPng Area, contd.

Simulation requirements
Routing Requirements
Market Viability
Transition Experiences
Transition Requirements
Accounting Requirements
Electric Power Research Comments
Cellular Industry View
Security Concerns
Italian Nuclear Physics Comments
Tactical Radio Requirements
Large Corporate Requirements
High Performance Networking Reqs.
ATM Support Requirements
Many Addresses per Host
Unix Host Requirements
Multiprotocol Interoperability

- Solicited IPng requirements outside IETF (RFC 1550)
  - Received 17 responses
    - RFCs 1667-1683
- Held IPng requirements BOF
  - Developed technical criteria RFC (RFC 1726)
    - Evaluated requirements submissions as part of determining criteria
    - Edited by Craig Partridge and Frank Kastenholz

8

Technical criteria


Craig Partridge


Frank Kastenholz

- Complete specification
- Architectural simplicity
- Scale
- Topological flexibility
- Performance
- Robust service
- Transition
- Media independence
- Datagram service
- Configuration ease

9

Technical criteria, contd.


Craig Partridge


Frank Kastenholz

- Security
- Unique names
- Access to standards
- Multicast support
- Extensibility
- Service classes
- Mobility
- Control Protocol
- Tunneling support

May 1994: IPng Area Directorate Retreat

- Evaluated proposals against criteria
- AD conclusion: none of the proposals met the criteria
- 2nd day: consolidated proposal
  - Good match to criteria
- Also developed proposal with variable length addresses
  - Failed to get IETF support

July 1994: IPng Decision

- Determined version number
  - Retrieved "6" from SIPP WG
    - "5" was assigned to Stream Protocol
- ADs presented IPng recommendation to IETF plenary in Toronto
  - Recommendation was for consolidated proposal published by SIPP WG
- Recommendation approved by IESG Nov. 17, 1994 (RFC 1752)

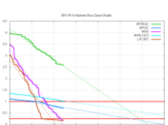
December 1995: IPv6 Specifications Published


Steve Deering


Bob Hinden

- RFC 1883: IPv6
  Steve Deering & Bob Hinden
- RFC 1884: IPv6 Addressing Architecture
- RFC 1885: ICMPv6
- RFC 1886: IPv6 DNS extensions
- RFC 1887: IPv6 Address Allocation
- RFC 1888: IPv6 and OSI NSAPs

Would we do it differently today?


IPv4 Address Runout

- CIDR & NATs pushed back the run-out time for v4 addresses a long time
  but have now actually run out
- Had we known the date at the time v6 was developed would the IETF have proceeded differently?
- First part of answer is to ask '*what did we do right and what did we do wrong*' using hindsight

What we did right (in my opinion)

- Reject circuits
- Simple protocol
- Lots of addresses
- Intrinsic security
  For some uses
- Lots of details
  neighbor discovery, auto configuration, link-local addresses, multiple addresses per interface, anycast, default router, no broadcast, simplifying router work
- Positioned for EID
  But that will likely never happen

What we did wrong (my opinion)

- Not enough different than IPv4
- Did not require host certificates
- Fixed length addresses
  - variable length would have been more future proof
- Interface addresses
  - Rather than "stack" - multiple addresses per host
    - See, for example, RFC 1681

What we did not do

Routing

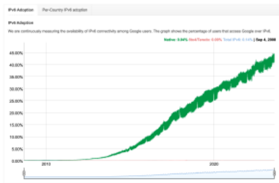
My hindsight

- Wrong to hurry (15 month process too short)
  - Tried to extend time but got too much pushback
  - Had time, since protocol basically defined in 1995
- Should have explored realities of performance impact of variable length addresses
- Wrong to punt on routing!

But

UPDATED

But result works and is being (slowly) deployed



Google: over 45% IPv6 usage – mid 2023
https://www.google.com/intl/en/ipv6/statistics.html#tab=ipv6-adoption

Image credits

Slide# credit
2 IETF logo: IETF
4 Frank Solensky - 1990
5 ROAD working group - 1992
6 Bradner photo: Harvard University Gazette
Mankin photo: https://www.verisigninc.com/en_US/innovation/verisign-labs/innovators/allison-mankin/index.xhtml
9 Partridge photo -http://www.ir.bbn.com/~craig/
Kastenholtz photo -https://www.linkedin.com/pub/frank-kastenholz/2/113/a08
10 Partridge photo - http://www.ir.bbn.com/~craig/
Kastenholtz photo - https://www.linkedin.com/pub/frank-kastenholz/2/113/a08
11 Logo: Big Ten Conference
12 IETF logo: IETF
Toronto logo: Toronto, Canada

Image credits

All drawings and photos by Scott Bradner unless noted
Slide# credit
13 Deering photo - cmu.edu
Hinden photo - Internet Society
14 Chart - potaroo.net
15 logo - IPv6 Forum
16 logo - IPv6 Forum
18 Bradner photo - Futureweb2010
19 both logos - Internet Society

Internet Protocol Suite

IPv6 Extension Headers

CSCI E 45a: The Cyber World – part A

1

IPv6 Extension Headers

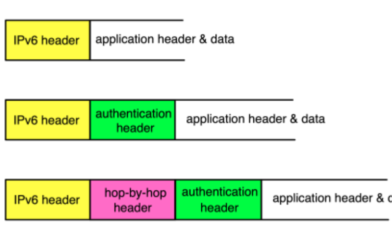
```
Hop-by-Hop Options
Header
Routing Header
Fragment Header
Authentication
Header
Encapsulating
Security Payload
Destination Options
Header
```

- Less used functions moved to extension headers
  - Only present when needed
- Only looked at by node with address in Destination Address field
  - Except Hop-by-Hop Options
  - Reduce router processing requirements
- Extensible

2

IPv6 Extension Headers, contd.

- Optional multiple headers



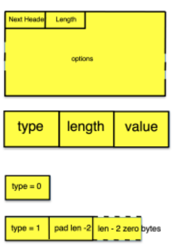
3

IPv6 options

- Some done with separate extensions
  - e.g., source route
- Other useful ones will be done within option headers
  - Hop-by-Hop Options Header
    - processed by all routers along path and by destination node
  - Destination Options Header
    - processed only by node(s) whose address(es) is(are) in destination address field

4

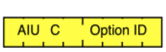
Hop-by-hop & Destination Options



- Option headers can contain multiple options
- Options in TLV format
  - Type-Length-Value
    - Type: identifies type of option
    - Length: option value field length
    - Value: option value
- Padded to 64-bit boundary
- Pad options

5

Type field: Header option handling



- AIU - action to be taken if option unknown by receiver
  - 00: skip this option
  - 01: discard the packet
  - 10: discard the packet & send ICMP error message
  - 11: 10 if not multicast destination
  - Eases introduction of new options
- C - set if option data can change en-route
  - (Hop-by-Hop Options Header only)
  - Say to include option in the authentication integrity assurance computation or not

6

Jumbogram Option

- If Payload Length field in IPv6 header = 0
  - Find actual payload length in jumbogram option in Hop-by-Hop Options Header
- Supports up to 4,294,967,296 byte (4 GB) packet length
  - Minimum value: 65,536
- Must not be used with Fragment Header

7

Image credits

All drawings by Scott Bradner

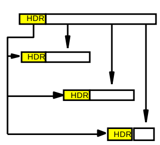
8

Internet Protocol Suite

Fragmentation

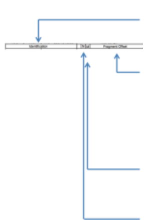
CSCI E 45a: The Cyber World – part A

Fragmentation



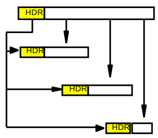
- Split packet into fragments if full size packet can not fit on output network
- Reassembly only done by destination node
- IPv4: source node and routers along the way can fragment
- IPv6: only source node can fragment
  - Reduce router processing load

Fragmentation fields



- 16-bit Identification field
  - Identify original packet
- 13-bit Fragmentation Offset field
  - Say where data was in original packet, 8 octet multiples
- 1-bit more-fragments field, 0 in last fragment sent
- 1-bit do-not-fragment field
  - Not present in IPv6

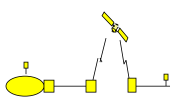
Fragmentation, process.



- Replicate IP header in each fragment
  - Modify fragmentation fields in IP header as needed
- Fragments must be reassembled by destination node
  - May be the only point in an arbitrary network to get all the fragments

4

Fragmentation, contd.



- Whole packet must resent if any fragments lost
  - ICMP time exceed message sent if host times out while rebuilding packet
- Min MTU
  - IPv4: 68 B
  - IPv6: 1280 bytes
- Min reassembly buffer
  - IPv4 576 B
  - IPv6 1500 B

5

Fragmentation: Path MTU

- Fragmentation hurts
  - Don't do it
- In theory, use Path MTU
  - Probe path to find largest packet that can reach destination
- Some problems
  - Paths need to be reasonably stable
  - Some black holes (e.g., firewalls block ICMP responses)
- If MTU of path shrinks
  - sender will receive ICMP Packet too Big message

6

Image credits

All drawings by Scott Bradner

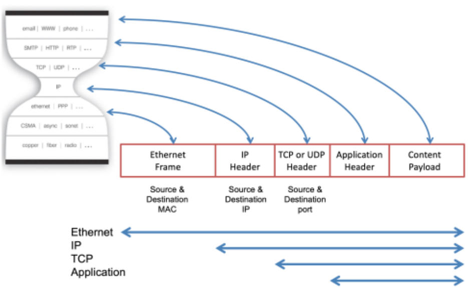
7

Internet Protocol Suite

Riding on IP

CSCI E 45a: The Cyber World – part A

Layered encapsulation



Riding on IP

UDP

ICMP

- User Datagram Protocol (UDP)
  Same semantics as IP (best effort delivery of datagrams)
- Internet Control Message Protocol (ICMP)
  Control, error and diagnostic messaging (unreliable)

Riding on IP, contd.

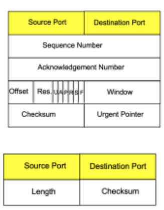
TCP

SCTP

QUIC

- Transmission Control Protocol (TCP)
  - Application to application reliable data stream
- Stream Control Transmission Protocol (SCTP)
  - Alternative to TCP, provides additional functions
- QUIC
  - UDP-based, stream-multiplexing, encrypted transport protocol
  - IETF revising Google proposal
- 100 or so other protocols

Ports

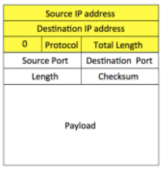


- UDP, TCP & SCTP include "ports"
  - Source port & destination port
- Ports used to multiplex & demultiplex packet streams to or from same node

Port, contd.

- Ports 1-1,023: "well known ports" registered with IANA
  - e.g., TCP port 25 is SMTP (email)
- Ports 1,024-49,151: other IANA-registered ports
  - often vendor specific
- Ports 49,152-65,536: dynamic
  - can not be registered

Pseudo header checksum



- Used by UDP & TCP
- Checksum calculated over the UDP or TCP part of the packet prepended with a "pseudo header" consisting of:
  - Source & destination IP addresses
  - Protocol field
  - Higher-level length field
- Including IP address fields detects miss-delivered packets

Image credits

All drawings and photos by Scott Bradner unless noted

Slide# credit

2 hourglass –Jonathan Zittrain - e.g. http://dltj.org/article/hourglass-national-e-book-program/

Internet Protocol Suite

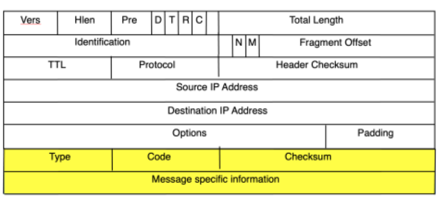
Internet Control Message Protocol (ICMP)

CSCI E 45a: The Cyber World – part A

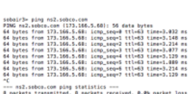
ICMP

0	Echo reply
3 unreachable	Destination
5	Redirect
8	Echo request
9	Router advertisement
10	Router solicitation
11	Time exceeded
12	Parameter problem
13	Timestamp
14	Timestamp reply

- Diagnostic or control messages
- Requesting information or reporting errors



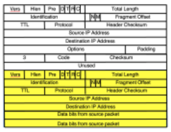
Ping (ICMP echo request) (Type = 8)



```
sebair3> ping ns2.sobco.com
PING ns2.sobco.com (173.166.5.68): 56 data bytes
64 bytes from 173.166.5.68: icmp_seq=0 ttl=63 time=3.032 ms
64 bytes from 173.166.5.68: icmp_seq=1 ttl=63 time=3.148 ms
64 bytes from 173.166.5.68: icmp_seq=2 ttl=63 time=3.214 ms
64 bytes from 173.166.5.68: icmp_seq=3 ttl=63 time=3.077 ms
64 bytes from 173.166.5.68: icmp_seq=4 ttl=63 time=3.129 ms
64 bytes from 173.166.5.68: icmp_seq=5 ttl=63 time=1.889 ms
64 bytes from 173.166.5.68: icmp_seq=6 ttl=63 time=3.214 ms
64 bytes from 173.166.5.68: icmp_seq=7 ttl=63 time=3.129 ms
^C
--- ns2.sobco.com ping statistics ---
8 packets transmitted, 8 packets received, 0.0% packet loss
round-trip min/avg/max/stddev = 1.889/2.979/3.214/0.416 ms
```

- Used for testing
  - Send ICMP echo requests to target host
    - With sequence numbers
  - Display each returned response
    - With round trip time
- Shows network reliability & latency

Destination unreachable (Type = 3)



- Returned by router if it receives a packer with an unreachable destination address
- Includes header and part of original packet

Code	
0	net unreachable
1	host unreachable
2	protocol unreachable
3	port unreachable
4	fragmentation required and DF set
5	source route failed

Image credits

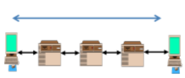
All drawings by Scott Bradner

Internet Protocol Suite
Flow & congestion control

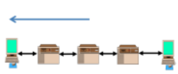
CSCI E 45a: The Cyber World – part A

1

Flow v. congestion Control



- Flow control ≠ congestion control
  - Flow control is end to end
    - Source waits for destination to say when it is ready for more data
  - Congestion control is middle to end



    - Network says when its overloaded (or about to be)
      - By losing or marking packets
    - Source slows down transmission rate

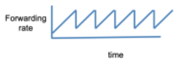
2

Fast computers & congestion

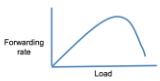
- Today's computers are almost always faster than the network
  - Thus, a single computer can often saturate its attached network link
- There may also be congestion on link to target computer if the target computer is engaged in multiple simultaneous sessions or its link is slower

3

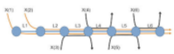
Congestion Control Goals



- Maximum rate of transfer for each session considering current network conditions
  - Respond to changes in network conditions

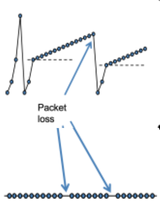


- Avoid congestion collapse
  - i.e., avoid multiple copies of a packet in transit



- Fair allocation of network capacity
  - At least between congestion responsive protocols

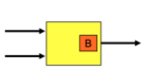
Congestion Responsive Protocols



- Congestion responsive protocols respond to changing network conditions
  - e.g., packet loss causes reduced transmission rate
- Congestion unresponsive protocols do not respond to changing network conditions
  - At least not quickly
    - Some applications have a slow feedback loop (e.g. RTCP)

Network Features

- A network consists of one or more interconnected network segments
- Network segments are interconnected with switches or routers

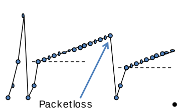


- Switches & routers include buffers
  - To deal with case of more data to send than output link can handle at any one instant

Buffering

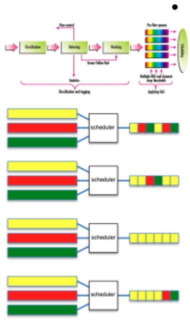
- Buffer used to smooth data flow to output
- Packet transmission rate depends on speed & load of output link
- Packets lost if buffer fills up
  - Called "tail drop" - last received packets are dropped
    - There are other options (e.g., active queue management) - will discuss later

Packet Loss



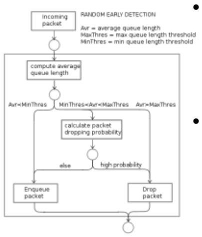
- Lost packets are used by end systems to indicate network congestion
  - i.e., more data in network than network can handle
    - Or at least more than the "bottleneck link" can handle
- Responsive protocols slow down transmission rate when packets are lost
- Wireless networks have non-congestion-based packet loss

Other Router-Based Mechanisms



- Per flow queuing
  - flow = communications session
    - Defined by "5-tuple" (source & dest ports & addresses + protocol)
- Transmission algorithms
  - Round-robin
    - Split link evenly between queues
  - Weighted round-robin
    - Split link based on some factor (e.g., customer link speed)
  - Priority
    - Higher priority traffic sent first
  - Controlled rate
    - e.g., fixed maximum rate

Router Queues



- Individual queues in router could be tail-drop or use active queue management (e.g., RED)
- Router could implement different quality of service mechanisms
  - Discussed in QoS lecture

10

Explicit Congestion Notification (ECN)


Sally Floyd

- TCP uses packet loss as for rate control
  - But data also lost - forcing retransmission
- ECN routers mark packets with CE flag if queue is more than a threshold full
- End systems treat CE-SEEN marked ACKs as lost data packets for rate control
  - But do not need to retransmit data

11

Image credits

All drawings and photos by Scott Bradner unless noted

Slide#	credit
3	congestion - congestion – http://indiatransportportal.com/continuous-digging-of-roads-leading-to-traffic-congestion-14832
4	fair allocation – http://scenic.princeton.edu/MRA/fairpracticenum.html
6	router icon -
8	WiFi logo – https://commons.wikimedia.org/wiki/File:Wi-Fi_Logo.svg
9	per flow queuing – http://m.eet.com/media/1100114/SS1140_MMC_PG_130.gif
10	RED diagram – https://en.wikipedia.org/?title=Random_early_detection
11	Floyd photo – https://http.icsi.berkeley.edu/icsi/people/floyd

12

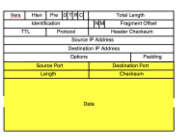
Internet Protocol Suite

User Datagram Protocol (UDP)

CSCI E 45a: The Cyber World – part A

1

UDP



- Try to transfer a “bundle of bits” to another host
- Best effort service
  - No flow control
  - No congestion control
  - No reliability check
  - No sequence check
- Basically IP + ports for multiplexing
  - (optional) “pseudo header checksum” that covers payload

2

UDP, uses

SNMP

DNS

- Voice over IP
- Streaming audio & video
- Network management
- Domain Name Service (DNS)
- Supporting new higher-level protocols

3

UDP, issues

- Non-responsive to congestion
  Can overwhelm TCP sessions
- Often blocked by firewalls

4

Image credits

All drawings and photos by Scott Bradner unless noted

Slide# credit

3 Vonage logo - http://www.vonage.com/
Apple FaceTime logo - https://www.apple.com/ios/facetime/

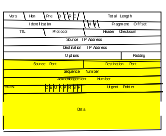
4 congestion - http://indiatransportportal.com/continuous-digging-of-roads-leading-to-traffic-congestion-14832

5

Internet Protocol Suite
Transmission Control Protocol (TCP)

CSCI E 45a: The Cyber World – part A

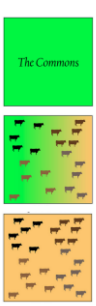
Transmission Control Protocol (TCP)



C Congestion Window Reduced (CWR)
E ECN Echo (ECE)
U Urgent Flag (URG)
A Acknowledgement (ACK)
P Push (PSH)
R Reset Connection (RST)
S Sync flag (SYN)
F Final Data (FIN)

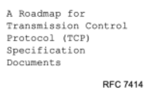
- Creates a reliable data stream between end hosts
  - Network is unaware of TCP
- Includes
  - Reliable start up & tear down
  - End-to-end flow control
  - Reacts to network conditions
- Deals with
  - Lost, corrupted or reordered packets

TCP Features



- Flow control in end systems
  - Reacts to available resources in end systems & to changing network conditions
- Control algorithms in the end systems must be compatible
  - No policing mechanisms to enforce compatibility
  - See *The Tragedy of the Commons*
    - Cheaters (as long as there are not too many) benefit from cheating
- New control algorithms need to understand impact on TCP

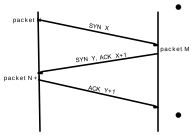
What is TCP?

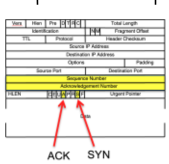


- Many updates to original RFC
- Not easy to tell what an implementer has to do
- IETF working group defined "what is TCP" - RFC 7414
  - Lists 128 normative RFCs
    - Published in February 2015, out of date about when it was published

4

TCP start up: 3-way handshake

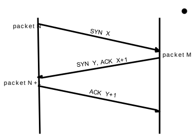




- Need to reliably establish state
  - state includes sequence numbers for each direction in each end
- Sequence:
  - I send you a start connection (SYN)
    - I include my sending seq number
  - You acknowledge my seq # (ACK)
    - Include your sending seq #
  - I acknowledge your seq #
- Result is state in both ends

5

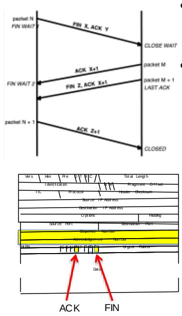
Why a 3-Way Handshake?



- You know you are talking with a real node
  - i.e. it responds
    - Routing infrastructure ensures packet went to the "right" place
  - 2-way could be spoofed
  - Man-in-the-middle can still be a problem
- Work reliably in the face of duplicate or lost packets
  - Not get hung with crashed end

6

TCP tear down: 3-way handshake



- 3-way handshake like startup
- must also wait 2 MSL before reusing the same address/port combination so connection is not confused

  MSL = maximum segment lifetime

  MSL = 2 minutes

7 Copyright © Scott Bradner & Ben Gaucherin 2015

SYN Attack

February 5-11, 2000

- Background
  - when SYN received
    - Store connection state in buffer
    - Send ACK
    - Wait for response - 75 sec timeout
    - Delete from buffer upon handshake or timeout
- Attack
  - Send many SYNs
    - From "random" source IP addresses
  - Buffer fills up
  - New connections never start

8 Copyright © Scott Bradner & Ben Gaucherin 2015

Cookie Defense

Dan Bernstein

- Forces client to keep & resend state
- When SYN received by server
  - Encrypt connection state
  - Send encrypted state (cookie) back to client in ACK
  - Forget about connection attempt
- Client
  - Include cookie in ACK of ACK
- Server
  - Decrypt and store connection state, start session

9 Copyright © Scott Bradner & Ben Gaucherin 2015

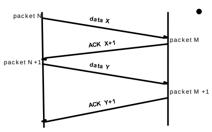
Initial Sequence Number

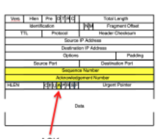
Robert T. Morris

- Original suggestion: use low-order 32 bits of 4 usec clock'
- Security issue - spoofing attack
  - *A Weakness in the 4.2BSD Unix TCP/IP Software* - R. T. Morris - 1985
- RFC 1948 describes problem and suggests alternate ways to create initial sequence number to avoid spoofing attack

10

Reliable Data Exchange

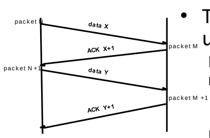




- Uses sequence numbers
  - Transmitter sequence number indicates last data byte it transmitted
  - Receiver sequence number indicates next data byte it expects
    - Thus acknowledging data up to that point

11

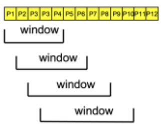
Reliable Data Exchange, contd.



- Timeout & duplicate ACKs used to identify lost packets
  - Retransmit if acknowledgement not received in time
    - Timeout value based on smoothed round trip time: min 1 sec
  - Lost and too-long-delayed packets are treated the same way
    - Could be lost data packet or lost ACK packet
- "Duplicate ACK" out of order packet, could mean loss
  - Reacknowledge previous packet

12

TCP flow control



- I send you some packets, you tell me when you are ready to accept more
- TCP uses a "window" to allow more than one packet in flight at same time
  - Sends an initial burst of packets (2-10)
- Acknowledgements (ACKs) of received packets authorize the sending of additional packets

TCP congestion response


Van Jacobson

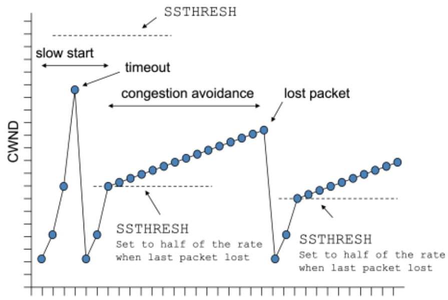
- Aims:
  - Maximize packet rate through network
    - But do not overload network
  - Share network fairly
- Modify window size to control transmission rate
  - Grow window if no congestion
  - Shrink window if congestion
- Congestion indicated by packet loss

TCP congestion response, phases

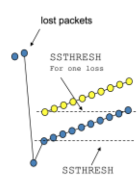

Van Jacobson

- Session startup: "slow start"
  - Rapidly determine rate where packet losses start
    - Window size doubled for each ACK
    - Until packet lost
- Session maintenance: "congestion avoidance"
  - Window size incremented by 1 packet for each ACK
    - Until packet lost
- Cut window size half for each packet loss & redo slow start

TCP timing

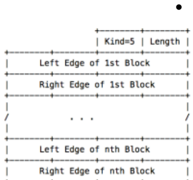


Multiple Packet Loss



- Multiple packet losses in same window or the loss of a retransmission is treated as multiple separate indications of congestion
- Thus cwnd (and ssthresh) MUST be lowered multiple times
- This is why active buffer management helps

Selective Acknowledgment (SACK)



RFC 2018

- RFC 2018
  - Defines TCP options that can be used to note missing data when data has been received after dropped packets
  - Sender figures out gap(s) & retransmits just the missing data from received SACK
  - Avoids unneeded retransmission and extra transmit rate back-off

Congestion collapse

- TCP uses additive-increase / multiplicative-decrease (AIMD)
  - Slow rate increase if no packet drops, fast rate decrease if packet dropped
- Failure to follow algorithm can result in congestion collapse
  - Very heavy network load, very long latency, very low throughput

ARPANET link LBL to UC Berkeley, 3 hops

Normal link throughput 32 Kbps

October 1986 40 bps

19 Copyright © Scott Bradner & Ben Gaucherin 2015

Buffer Bloat

Jim Gettys

- Buffers in many network devices now far too big
  - Memory too cheap
- Buffers fill up and stay full
- Can add seconds of latency
- At all levels of network
  - from LAN drivers to backbone routers

20 Copyright © Scott Bradner & Ben Gaucherin 2015

TCP security

Robert T. Morris

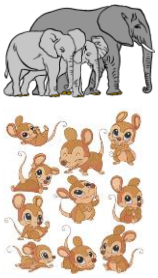
- TCP security ensured by
  - Routing infrastructure
  - Lack of knowledge of sequence number in correct range
    - Vulnerability if sequence number can be guessed
  - Crypto checksum option
    - RFC 5925

```
+------------+------------+------------+------------+
|  Kind=29   |   Length   |   KeyID    | RNextKeyID |
+------------+------------+------------+------------+
|                     MAC           ...
+-----------------------------------...

   ...-----------------+
   ...  MAC (con't)    |
   ...-----------------+
```

21 Copyright © Scott Bradner & Ben Gaucherin 2015

TCP Issues



- Elephants vs. mice
- Many concurrent flows between same hosts
- Large bandwidth / delay products
  - e.g. satellites
- Non-congestion-based packet loss (e.g. wireless)
- TCP spoofers
  - Fiddle with TCP flows to try and control them

22

Image credits

All drawings and photos by Scott Bradner unless noted

Slide#	credit
3	all 3 drawings - Stephens Planning & Design
8	CNN logo – www.cnn.com
	ebay logo – www.ebay.com
	Yahoo! logo – www.yahoo.com
	Amazon logo – www.amazon.com
9	Bernstein photo - https://en.wikipedia.org/wiki/Daniel_J._Bernstein
14	Jacobson photo - http://www.pcmag.com/slideshow_viewer/0,3253,l=209433&a=209433&po=9,00.asp
18	Section 3 RFC 2018 – www.ietf.org/rfc/rfc2018.txt
20	Gettys photo - https://en.wikipedia.org/wiki/Jim_Gettys
21	Figure 2 RFC 5921 – www.ietf.org/rfc/rfc5921.txt
22	Elephant clip art - clipartpanda.com
	mice clip art - embroiderypassbook.com

23

Internet Protocol Suite

Stream Control Transmission Protocol (SCTP)

CSCI E 45a: The Cyber World – part A

1

SCTP

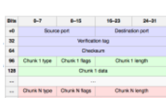
Randall Stewart

Qiaobing Xie

- Originally designed to support telephone signaling over the Internet & to run over UDP
  - IETF sigtran working group
  - Required low latency and reliability
- IETF Transport ADs asked authors to redesign it to run over IP and be a "TCPng"
  - What TCP would look like if it were redone

2

SCTP, contd.



Bits	0–7	8–15	16–23	24–31
+0	Source port		Destination port	
32	Verification tag			
64	Checksum			
96	Chunk 1 type	Chunk 1 flags	Chunk 1 length	
128	Chunk 1 data			
...	...			
...	Chunk N type	Chunk N flags	Chunk N length	
...	Chunk N data			

- TCP compatible congestion control
- Multi-stream
- Message-framing
  - Rather than stream, like TCP
- Supports multi-homing
- Can support unordered delivery
- Stateless session startup
  - Cookie-based

3

SCTP, contd.

WebRTC

- Used over UDP in WEB RTC (Real Time Collaboration on the World Wide Web)

Image credits

All drawings and photos by Scott Bradner unless noted

Slide# credit

2 Stewart photo - http://people.freebsd.org/~rrs/

Xie photo - http://www.prnewswire.com/news-releases/adara-networks-appoints-dr-qiaobing-xie-chief-technologist-300033692.html

3 diagram - https://en.wikipedia.org/wiki/Stream_Control_Transmission_Protocol

multi-homing diagram - http://www.ibm.com/developerworks/library/l-sctp/

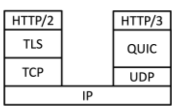
4 webrtc logo - http://www.nethram.com/webrtc-with-asterisk-12/

Internet Protocol Suite
QUIC

CSCI E 45a: The Cyber World – part A

1

QUIC

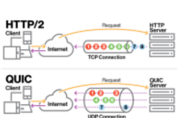


RFC 8999
RFC 9000
RFC 9001
RFC 9002

- Original idea by Google
  - used to speed up web traffic
- Google offered it to the IETF
- Evolved into TCP alternative
  - not limited to HTTP
- Sort of a TCP-SCTP-TLS mashup over UDP
  - over UDP so it can be deployed
  - also, can be run at user level
- Note: QUIC is a name not an acronym

2

QUIC, contd.



- Establishes secure connection between Internet nodes
- One or more independent streams run in connection
- Connection-level TCP compatible congestion control module
  - can be replaced
- Reliability & flow control at stream level
- Uses port 443 for web traffic

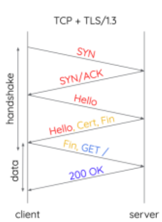
3

QUIC, Connection IDs

```
Initial Packet {
    Header Form (1) = 1,
    Fixed Bit (1) = 1,
    Long Packet Type (2) = 0,
    Reserved Bits (2),
    Packet Number Length (2),
    Version (32),
    Destination Connection ID Length (8),
    Destination Connection ID (0..160),
    Source Connection ID Length (8),
    Source Connection ID (0..160),
    Token Length (i),
    Token (..),
    Length (i),
    Packet Number (8..32),
    Packet Payload (8..),
  }
```

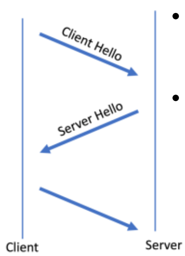
- Connections identified by unencrypted IDs
- IDs can be used by load balancers
- Can isolate connection from underlying addressing
  - connection can migrate between IP addresses
    - e.g., cellular to WiFi

QUIC, Connection Setup



- Uses TLS 1.3 handshake
  - all packets encrypted
- Minimize round trips needed to set up connection & start sending data
  - TCP/TLS can take 3-4 RT
  - QUIC 1-RTT – data after 1 RTT
  - QUIC 0-RTT – data with 1st packet
    - uses server info cashed by client
- Uses "key share"
  - contains info for Diffie-Hellman key agreement mechanism

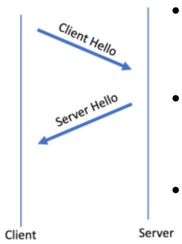
QUIC, TLS 1.3: New Server (1-RTT)



bold: encrypted

- Client: Client Hello
  - {supported ciphers, key agreement mode, key share}
- Server: Server Hello
  - {chosen cipher, key agreement mode, key share, pre_shared key, **cert, signature, server finished**}
    - signature covers both client & server hellos
- Client:
  - {**client finished, [data]**}

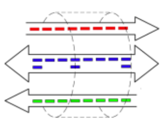
QUIC, TLS 1.3: Known Server (0-RTT)



- Client: Client Hello
  key share, key agreement mode, pre shared key, **data**
- Server: Server Hello
  key share, pre_shared key, **server finished, data**
- Vulnerable to replay attack

bold: encrypted

QUIC, Streams



- Multiple streams per connection
  Each has a 62-bit ID
- Uni- or bi-directional
- Streams are independent
  do not block other streams
    no TCP "head of line blocking"
- Can have packets from multiple streams in same connection frame
  Datagrams can be data or control

QUIC, HTTP/3

faster & more secure

- HTTP/3 optimized to run over QUIC
- Uses QUIC streams instead of multiple TCP connections
  no head of line blocking
- Always encrypted (HTTPS)
  Blocks middlebox manipulation

QUIC, Future

- Forward Error Correction
- Different congestion control modules
- Version negotiation
- More than just the web
- Datagrams

Image credits

Drawings and photos by Scott Bradner unless noted

Slide# credit

2 https://daniel.haxx.se/blog/tag/ietf/

3 https://www.verizondigitalmedia.com/blog/how-quic-speeds-up-all-web-applications/

4 https://www.rfc-editor.org/info/rfc9000

5 https://www.semanticscholar.org/paper/Analysis-of-QUIC-Session-Establishment-and-Its-Gagliardi-Levillain/dca3f6733638076020af5a32f3fe8e9f23912916

10 https://www.ideal-ist.eu/spotlight/innovations-breakthroughs-what-future-will-bring-us

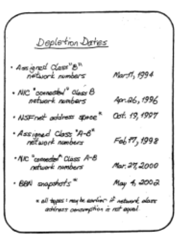
Internet Protocol Suite

Conclusion

CSCI E 45a: The Cyber World – part A

1

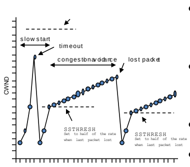
IP as internetworking bearer service



- Internetwork addresses support global networks
- The Internet Protocol provides a basic transport service
- IPv6 was necessitated by running out of IPv4 addresses

2

IP as support



- Higher-level protocols use IP for transport
- UDP provides a simple non-corrected datagram service
- ICMP is used for signaling
- TCP provides a reliable data stream that reacts to network and host resources
- QUIC provides secure TCP alternative

3

Image credits

All drawings and photos by Scott Bradner unless noted

Slide#	credit
2	Frank Solensky - 1990

4